

DOCUMENT RESUME

ED 155 198

TM 007 099

AUTHOR Malbon, Scott
TITLE Learning Disability Screening Instruments: Discriminative Validity.
PUB DATE Mar 78
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, March 27-31, 1978)
EDRS PRICE MF-\$0.83 HC-\$1.67 Plus Postage.
DESCRIPTORS Behavior Rating Scales; Cutting Scores; Elementary Education; *Learning Disabilities; *Predictive Validity; *Screening Tests; Special Education; Student Placement; Testing; Testing Problems; *Test Selection; *Test Validity
IDENTIFIERS Pupil Rating Scale

ABSTRACT

Screening refers here to the process of making an initial selection of those who need further diagnostic evaluations. A screening process which effectively selects true learning disability (LD) cases from a population reduces the number of incorrectly selected non-LD children who would be needlessly given costly psychoeducational evaluations. Although there are several LD screening instruments in wide use, empirical evidence of the validity of these instruments is sparse and often misdirected. Discussion of the merits of these instruments typically focuses on the relation of the screening device to the construct of learning disabilities. Instruments which have most successfully emulated the components of a full diagnosis receive favorable evaluations. Although this is probably a wise approach in instrument development it is not the most important consideration in the evaluation of the effectiveness of the screening instrument in actual use. The empirical validity of a screening instrument is a function of correct classification, and this validity can be demonstrated by a simple 2 x 2 classification matrix. Validity data on Myklebust's Pupil Rating Scale is re-analyzed to demonstrate that high content validity does not automatically provide high discriminative validity. (Author/BW)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED155198

TM007 099

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

LEARNING DISABILITY SCREENING INSTRUMENTS:
DISCRIMINATIVE VALIDITY

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

Scott Malbon

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC) AND
USERS OF THE ERIC SYSTEM."

Scott Malbon

The Southwest Georgia Program
for Exceptional Children

Presented at the 1978 Annual Meeting of the American Educational Research
Association, Toronto.

Section 1.11

Reproduced in the U.S.A.

Learning Disability Screening Instruments:
Discriminative Validity

Scott Malton

The Southwest Georgia Program
for Exceptional Children

An important step in the delivery of special education to handicapped children is the identification of those eligible for services. For learning disability (LD) programs the psychoeducational evaluation necessary to place (label) children under recent state and federal regulations (P.L. 94-142) is very costly in terms of staff resources. A valuable asset to a LD program would be a screening instrument which could be administered quickly and cheaply to the school population, or to the large number of referred students, and which would indicate all LD students needing a full evaluation while diverting non-LD students from this laborious assessment and placement process.

Screening refers here to the process of making an initial selection of those who will receive diagnostic evaluations. This is a "pass-through" step the function of which is to reduce the number of students who must be given costly evaluations, thus maximizing staff and program resources, and also facilitating the lengthy placement process.

The explication of a precise definition reflecting the use and goals of the screening procedure is important because the intended use of the screening instrument determines which procedural constraints apply to the procedure and the type of validity data relevant to the evaluation of the instrument. The term screening is frequently used in two different ways in education literature: (1) as a tentative categorization (sorting) on the

basis of minimal information, and (2) early identification for the purpose of prevention. In this paper screening is used in the first sense. However, the current popularity of mass screening procedures in health and education for the purpose of identifying and treating mild problems before they become serious has given us the second, slightly different use of the term: early identification for the purpose of prevention. The definition intended is not always stated explicitly and sometimes the term is not used consistently. The definition issue is important because the intended use of the screening instrument determines the type of validity data which is relevant to the evaluation of the instrument. At a recent national convention we were present at a paper presentation ostensibly on screening in which the author used the term indiscriminantly to refer to screening (as we have defined it), psychological diagnosis, and educational assessment.

There are many instruments which could be applied to the screening task (Wallace and McLoughlin, 1975, pp. 292-295) including several instruments now available commercially which were developed specifically to serve the screening function. For the LD program director trying to choose a screening instrument the most logical and best single reference, is the instrument manual. Typically the manual provides a description of, and rationale for, the instrument, administration procedures, scoring, interpretation of scores, and developmental research. The research section reports procedures used in instrument development, reliability, and validity studies. Three types of research are generally reported which bear on the issue of validity: expert review, group difference studies, and correlation studies.

The expert review is generally presented as support for the instrument's content validity. These reviewers are sometimes paid consultants involved in the development of the instrument, or they are professionals asked to evaluate whether or not the procedures and items in the test will effectively identify LD students. However, lack of agreement about what constitutes, or indicates, a learning disability is the usual outcome of such surveys. These fundamental disagreements are usually side-stepped in the test manual (eg. Meier, et. al., 1970) but certainly should not be taken to indicate good content validity.

Another method of demonstrating the instrument's content validity is a review of the literature, citing publications (not always empirical studies) as support. Although these reviews are necessarily selective they avoid the lack of consensus obtained by the expert review technique.

With the emphasis on the relation of the screening device to the author's concept of learning disabilities, the validity presentations argue the merits of the screening instrument. Instrument evaluations of this type have a distinctly "clinical" perspective with the most favorable evaluations going to the instrument which has best emulated the components of a full diagnostic evaluation. Although this is probably a wise approach in instrument development, and is intuitively appealing to test consumers, it is not the most important consideration in the evaluation of the effectiveness of the screening instrument in actual use.

The heavy reliance on face validity and content validity is widespread. This is particularly the case when school systems or LD programs develop their own screening instruments and apply them with no empirical

evidence of reliability or validity. Intuitive clinical evaluation of the instrument is the most common, persuasive method of choosing a screening instrument.¹

The virtues of screening instruments, ease of administration chief among them, is obtained at the cost of reliability and validity - a trade-off familiar to educational researchers and program evaluators. However, the limitations inherent in screening instruments are not understood by teachers and administrators and misuse of screening instruments is common.

One misuse of screening procedures is based on their rather substantial face validity. The favorable intuitive clinical evaluation of the instrument leads to de facto diagnosis of learning disabilities. ("Oh, this test shows that Johnny has an auditory processing deficit.") This, in turn, leads to de facto labeling and the potentially serious consequences which may follow.

One of the most serious of consequences is placement in special education classes. Although I have no direct evidence of students being placed in special education classes on the basis of screening instruments alone, I would not be surprised to see that. I do know of students in large metropolitan school systems being placed in special education LD classes without psychological evaluations.

Screening procedures may be given to entire classes, or school populations without singling out individual children; results are not used to label or place children. As a routine procedure, therefore, parental consent and due process requirements do not apply (this is one of the procedural conveniences of such instruments). However, if screening results are

misused to label and/or place students outside of due process requirements the misuse extends from measurement concerns to legal issues. The desire to avoid the multitude of due process requirements may produce pressures to rely more on screening instruments.

The misuse of screening instruments is due in part to an inadequate understanding of instrument validity, and test publishers do not help the situation. When validity research is present in test manuals most of it is directed towards face, content, and construct validity. In actual use a valid screening instrument is one which discriminates the target population (in this case students with learning disabilities) from the larger population. The simplest way to demonstrate this is a 2 X 2 classification matrix (See Figure 1) which illustrates how the binomial classification by the screening device agrees with the classification by the criterion measure(s). An instrument's high content validity and construct validity are relevant only if the instrument's discriminative ability is high as a consequence, which may, or may not, be the situation empirically. This discriminative validity may be viewed as predictive validity where subsequent (though immediate) classification by a full diagnostic evaluation is being predicted. Or equivalently, discriminative ability may be viewed as concurrent validity where agreement with a criterion measure is expressed in terms of agreement in a binary classification rather than a linear correlation.

Insert Figure 1 About Here

However, the program administrator today must choose a screening instrument without the relevant validity data; no screening instrument

publisher reports this data and to my knowledge there is no published research on discriminative validity of LD screening instruments.² In addition to this "sin of omission" test producers make a "sin of commission" when they claim discriminative ability on the basis of inadequate, or irrelevant, data.

When empirical evidence of instrument validity is collected correlational designs are frequently used. The screening instrument is administered, along with one or more highly regarded assessment instrument to a sample school population. Statistically significant correlations are taken to indicate the validity of the screening instrument. This is, indeed, evidence relevant to the instrument's concurrent validity. However, test producers erroneously infer from significant correlations that the screening instrument accurately identifies, or discriminates, LD children. This would be a valid argument only if the criterion instrument is regarded as sufficient to determine the presence of a learning disability in a student and if the correlation with the screening instrument was 1.00. To the extent that the correlations are less than 1.00 the screening instrument will misclassify students. (This is illustrated in the second part of this paper.)

A common study undertaken to demonstrate the validity of an instrument is to give the instrument to identified normal and LD groups, looking for significantly different group means. Although this procedure has been suggested for investigations of construct validity (Helmstadter, 1964) this is a weak test of the screening instrument's discriminative validity. Typically the score distributions of the two groups overlap considerably and statistically significant mean differences do not tell us how many

individuals in the two groups would be correctly classified by the screening process.

Reanalysis of the Pupil Rating Scale

In order to illustrate the fallacy of claiming discriminative validity on the basis of significant group differences data from the Pupil Rating Scale: Screening for Learning Disabilities (Myklebust, 1971) are presented and reanalyzed. The Pupil Rating Scale (PRS), one of the first produced commercially, is chosen because it is widely used and has received high acclaim. A review of the PRS in the Journal of Special Education stated: "Myklebust's scale is impressive. Because of its ease of usage and interpretation, it can be said to be psychometrically eloquent, a term rarely applicable to any measurement device! (Proger, 1973, p. 317)" In addition the PRS has received more, and better, research than most other screening instruments, and reports enough data to allow the reanalysis.

Description of the PRS

The PRS is attractive to the program administrator because it can be completed by the classroom teacher in approximately fifteen minutes per student. The PRS solicits judgements on 24 behavioral characteristics divided unequally into five areas: Auditory Comprehension (4), Spoken Language (5), Orientation (4), Motor Coordination (3), and Personal-Social Behavior (8). The Auditory Comprehension area, for example, contains the following behavioral characteristics: Comprehending word meaning, Following instructions, Comprehending class discussions, Retaining information. The teacher rates the student on each of the behavioral characteristics using a five-point scale where one represents below average

performance, three represents average performance, and five represents above average performance. For each behavioral characteristic, descriptive statements are provided to illustrate the performance level corresponding to each rating. For example, the descriptive statements corresponding to the five ratings for Comprehending word meaning are:

- 1 - Extremely immature level of understanding
- 2 - Fails to grasp simple word meanings; misunderstands words at grade level
- 3 - Good grasp of vocabulary for age and grade
- 4 - Understands all grade-level vocabulary as well as higher-level word meanings
- 5 - Superior understanding of vocabulary; understands many abstract words

The score for each behavioral area is simply the sum of the ratings for that area. In addition to a total score the sum of the first two behavioral areas is presented as a "Verbal Score" and the sum of the remaining three behavioral areas is presented as a "Nonverbal Score".

Interpretation of the obtained scores is inhibited by the different number of items in each area, and subtotals, and by the absence of tables to convert raw scores to standard scores. Descriptive statistics (means, standard deviations, and ranges) are presented for the research sample. However the sample consisted of third and fourth graders only from four metropolitan school districts and therefore lack of representativeness in different populations is a potential problem.

PRS Research

A great deal of research is reported in the PRS manual although much is of little use in the assessment of instrument validity. In place of

the parade of expert consultants who normally attest to the content validity of the instrument, author Myklebust relies on his own considerable stature and judicious use of references to the professional literature in the areas of concern. There is no explicit presentation of reliability data or construct validity although data relevant to that issue can be teased from several presentations.

The basis for the reported empirical research is a study of mean scores obtained by normal and learning handicapped students. Each student in the research sample, all of the third and fourth graders in four suburban school systems (N = 2176) was rated by his teacher on the PRS; other tests were administered by the research staff. The criterion test battery was used to determine each student's Learning Quotient (Myklebust, 1968)³ on the basis of which the students were categorized into two groups; Normal if the LQ was 90 or above, and Learning Problem if the LQ was 89 or below. The mean PRS scores for the two groups were significantly different, leading the author to conclude "Thus, it is clear that the Scale can be used to differentiate normal children from those with deficits in learning. (Myklebust, 1971, p. 12)"

From significant group differences one may assume the instrument to have some construct validity. However, these data do not warrant a claim of discriminative validity. In order to illustrate this erroneous inference we must look at the score distributions and determine how the individual students would be classified by the PRS.

Estimating the Discriminative Validity of the PRS

The distribution of the PRS scores for the two groups can be estimated

by assuming normal distributions and by using the reported Ns, means, and standard deviations. We chose a PRS score of 69 as the criterion score⁴; students with PRS scores of 69 or below are classified as having learning problems, students with scores of 70 or above are classified as normal. By placing a line at that point on the joint distribution (Figure 2) we see that many students are correctly classified but many have been misclassified in terms of the LQ criterion. The students in sections B and D of Figure 2 are correctly classified ("hits"); in area B those who were identified by the criterion as having learning problems were also so identified by the PRS ("true positives"), in area D those who were identified by the criterion as being normal were so classified by the PRS ("true negatives"). Students in sections A and C were incorrectly classified ("misses"); those in area A were identified by the criterion as normal were incorrectly classified as LD by the PRS ("false positive"), those in area C were identified by the criterion as LD but were incorrectly classified as normal by the PRS ("false negative"). This is represented in a 2 X 2 classification matrix (Figure 3).

Insert Figure 2 and 3 about here

The "hit rate" in this example is 73%; $(190 + 1396)/2176$. However, only 190/339, or 56% of the students with identified learning problems were correctly screened, and 441 identified normal students were screened as having learning deficits. These two errors are of concern to the LD program staff. The first problem is serious because children with learning problems (in this example, 149 students) would be diverted from further diagnostic evaluation and would not receive needed special education. The

second type of error, allowing non-handicapped children to pass through and receive full evaluations leads to inefficient use of staff; 441 of 631 of those sent on for testing will not be placed in the special program.

With any particular screening instrument the serious problem of false negatives can be reduced by increasing the criterion score on the screening instrument. In this example, if the cut-off is raised to 80 the false negatives decrease to 62. However, as a consequence of moving the cut-off score there is an increase in the false positives to 918 (Figure 4). This means that in order to correctly classify more of the actual learning problem children with a particular screening instrument by moving the cutting score, more normal children will be incorrectly classified as LD, thereby reducing the efficiency of the screening instrument. The final placement of the criterion score is usually determined jointly by the desire to identify as many LD children as possible, but within realistic resource constraints.

Insert Figure 4 about here

These error rates are probably not unusual for a quickly administered screening test. However, the teachers and administrators that we have questioned appear to be unaware of the rate, and type of misclassification in screening instruments and attribute more discriminative validity to them than is warranted empirically. This is probably a consequence of the high face validity given the instruments by test producers.

To illustrate the fallacy of inferring discriminative validity from significant, but less than perfect, correlations we must again examine

score distributions, in this case a scatter plot of the relationship of the two instruments (Figure 5). By placing lines on the scatter plot corresponding to the cutting scores for the criterion and screening measures we can observe how the screening instrument correctly and incorrectly classified students. Using our previous notation, areas B and D represent false positives and false negatives, respectively. Again the joint distribution can be represented in a 2 X 2 classification matrix.

Insert Figure 5 About Here

Incorrect classifications are a consequence of the modest correlation of the screening instrument and the criterion instrument. If the correlation was 1.00 (or very close) and the optimal screening score was applied, the hit rate would be 100% (Figure 6). However, such high correlations are rarely found in practice.

Insert Figure 6 About Here

Conclusions and Recommendations

Although screening instruments can be a valuable asset to LD programs there is potential for misuse. The misuse of screening instruments is a consequence of the lack of appropriate validity research conducted by test producers and the mistaken belief by program staff that good face and content validity guarantees good discriminative validity. In order to increase the appropriate use of screening instruments educational researchers and program evaluators should:

- (1) clarify the definition of the term screening, pointing out ambiguous and incorrect uses of the term.
- (2) remind program staff that existing screening instruments have modest, or unknown, reliability and validity. As in any decision-making setting the finality and weight of the decision should be warranted by the reliability and validity of the data upon which the decision is based.
- (3) discuss the validity of relevance for screening instruments.
Test producers should be encouraged to conduct relevant research and report findings straightforwardly; we recommend the simple 2 X 2 classification matrix. Users of screening instruments should be educated with respect to: (a) the various types of validity, (b) the relevance of discriminative validity for screening procedures, and (c) the error of inferring discriminative validity from face, content, or construct validity. The demonstration of score distributions as presented here may be an effective educational practice.
- (4) consider formal action on the establishment of standards for screening instruments. This may take the form of a formal recommendation, resolution, or report from professional organizations such as the AERA.

Footnotes

- 1 At a recent national conference an author of a home-made LD screening device, a 16 item checklist, was asked if the device has been administered to any LD children. The author responded that 12 of 14 identified LD children in a special class had been identified by the checklist. The author, with scornful sarcasm, then implied that the psychological evaluation upon which the other two students had been placed had obviously been in error; many in the audience nodded in agreement.
- 2 The author, at this writing, is aware of only one unpublished study (Donah, et. al., 1976) which presents the type of validity study of LD screening instruments advocated here.
- 3
$$LQ = \frac{\text{Achievement age}}{\text{Expectancy age}}, \text{ where Expectancy age} = \frac{(MA + CA + GA)}{3}$$
- 4 The PRS manual does not suggest a particular cutting score. Sixty nine is three points below the theoretical average and about nine points (and one-half standard deviation below) the average obtained in the research sample. The consequence of using other cutting scores is demonstrated later in the paper.

References

- Donah, C., Dinero, T., and Larson, G. Concurrent, Construct, and Predictive Validation of the Slingerland Screening Tests for Identifying Children with Specific Language Disability. Paper presented at the Annual Meeting of AERA, San Francisco, April 1976.
- Helmstadter, G. Principles of Psychological Measurement New York: Appleton-Century-Crofts, 1964.
- Meier, J., Cazier, V., Giles, M. Individual Learning Disabilities Classroom Screening Instrument: Administration and Scoring Manual with Remedial Suggestions. Evergreen Colorado: Learning Pathways, 1970.
- Myklebust, H. Definition and Overview. In H. Myklebust (Ed.), Progress in Learning Disabilities, Volume 1, New York: Gruen and Stratton, 1968.
- Myklebust, H. The Pupil Rating Scale: Screening for Learning Disabilities. New York: Gruen and Stratton, 1971.
- Proger, B. Test Review No. 15: The Pupil Rating Scale: Screening for Learning Disabilities. Journal of Special Education, 1973, 7 (3), 311-317.
- Wallace, G., and McLaughlin, J. Learning Disabilities: Concepts and Characteristics. Columbus, Ohio: Charles E. Merrill, 1975.

		Criterion Measure		
		LD	Normal	
Screening Instrument	LD	B True Positives	A False Positives	
	Normal	C False Negatives	D True Negatives	

"Hits" = B + D

"Misses" = A + C

Figure 1. Model Joint Classification Matrix

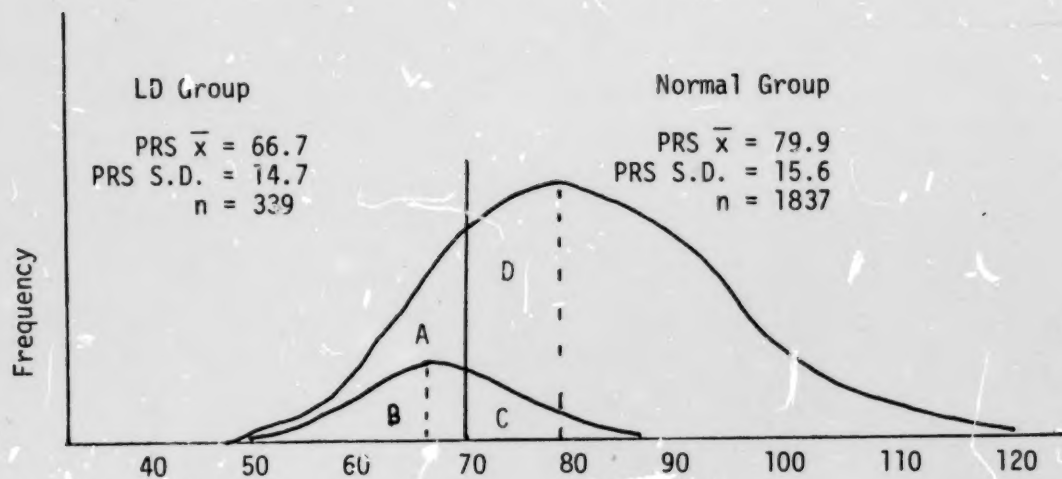


Figure 2. Distribution of PRS screening scores for both the identified LD and Normal groups.

		Criterion Measures		
		LD	Normal	
PRS (69)	LD	190	441	631
	Normal	149	1396	1545
		339	1837	

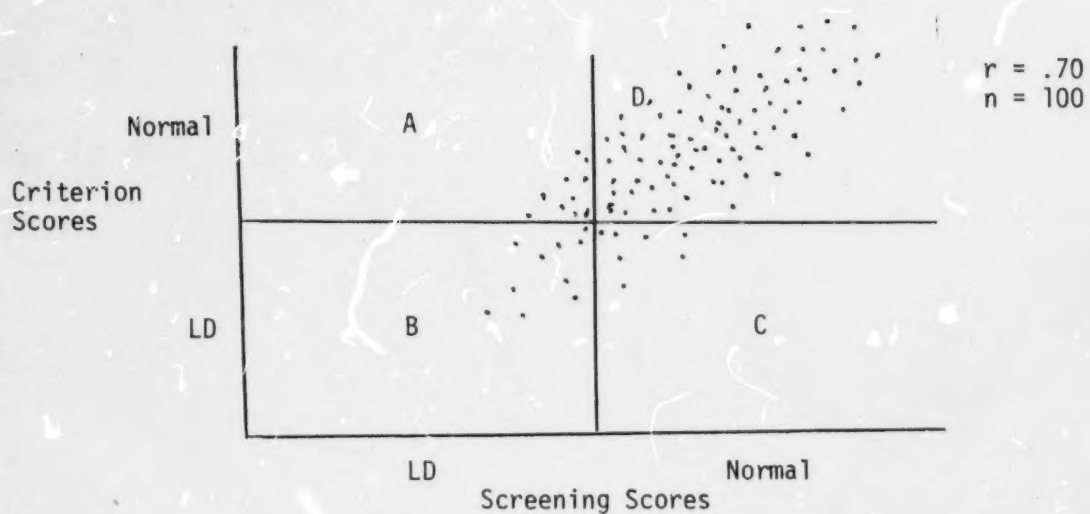
Figure 3. Classification Matrix for PRS data when 69 used as cut-off.

		Criterion	
		LD	Normal
PRS (62)	LD	127	228
	Normal	212	1609

		Criterion	
		LD	Normal
PRS (69)	LD	190	441
	Normal	149	1396

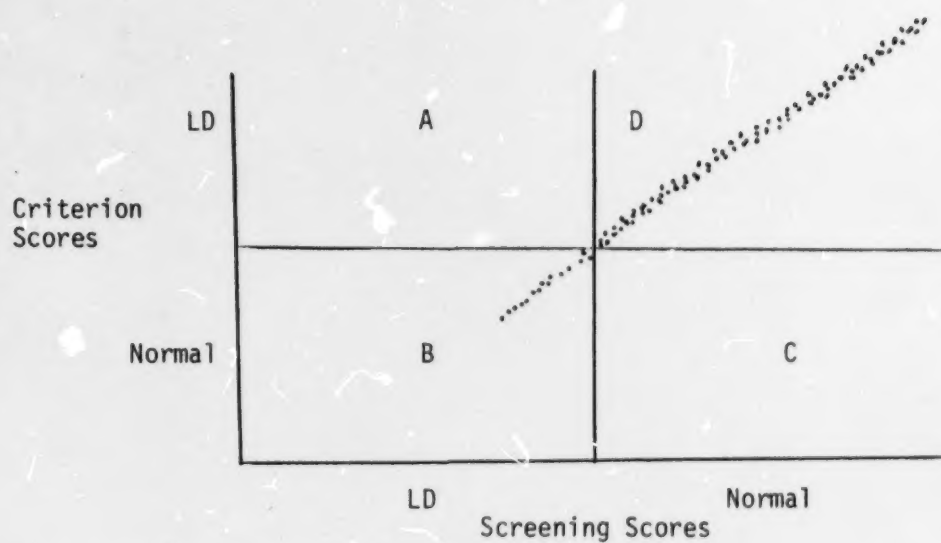
		Criterion	
		LD	Normal
PRS (80)	LD	277	919
	Normal	62	918

Figure 4. Series of Classification Matrices indicating the effect of applying different cut-off scores to the PRS data.



		Criterion Classification	
		LD	Normal
Screening Classification	LD	9	14
	Normal	6	71

Figure 5. Hypothetical scatter plot and corresponding Classification matrix for a Screening-Criterion correlation of .70 with 100 students.



		Criterion Classification	
		LD	Normal
Screening Classification	LD	15	0
	Normal	0	85

Figure 6. Scatter plot and corresponding classification matrix for a Screening-Criterion correlation of 1.00 with 100 students.